Handling the heterogeneity of genomic and metabolic networks data within flexible workflows with the PADMet toolbox

Marie Chevallier, Méziane Aite, Jeanne Got, Guillaume Collet, Nicolas Loira, Maria-Paz Cortés, Clémence Frioux, Julie Laniau, Camille Trottier, Alejandro Maass, et al.

Handling the heterogeneity of genomic and metabolic networks data within flexible workflows with the PADMet toolbox

Marie Chevallier*†[1], Meziane Aite[1], Jeanne Got[1], Guillaume Collet[1], Nicolas Loira[2], Maria Paz Cortes[2], Clémence Frioux[1], Julie Laniau[1], Camille Trottier[1], Alejandro Maas[2], and Anne Siegel‡[1]

[1]DYLISS (INRIA - IRISA) – INRIA, Universite de Rennes 1, CNRS : UMR6074 – Campus de Beaulieu 35042 Rennes cedex, France

[2]Centre de Modélisation Mathématique / Centro de Modelamiento Matemático (CMM) – Chili

Résumé

A main challenge of the era of fast and massive genome sequencing is to transform sequences into biological knowledge. The reconstruction of metabolic networks that include all biochemical reactions of a cell is a way to understand physiology interactions from genomic data. In 2010, Thiele and Palsson described a general protocol enabling the reconstruction of high-quality metabolic networks. Since then several approaches have been implemented for this purpose, such as Model Seed (*Henry et al., 2010*) , the Cobra ToolBox (*Becker et al., 2007*) and the Raven Toolbox (*Agren et al., 2013*). These methods rely mainly on drafting a first metabolic network from genome annotations and orthology information followed by a gap-filling step. More precisely, in the case of exotic species the lack of good annotations and poor biological information result in incomplete networks. Reference databases of metabolic reactions guide the filling process in order to check whether adding reactions to a network allows compounds of interest to be produced from a given growth media. As a final objective, as soon as the network is considered to be complete enough, functional studies are undergone, often relying on the constraint-based paradigm derived from the Flux Balance Analysis (FBA) framework (*Orth et al., 2010*).

The high diversity of input files and tools required to run any metabolic networks reconstruction protocol represents an important drawback. Genomic data is often required, provided in different formats: either annotated genomes, and/or protein sequences, possibly associated with trained Hidden Markov Models. In addition, most approaches require reference metabolic networks of a template organism. Dictionaries mapping the reference metabolic databases to the gene identifiers corresponding to the studied organism may be required. As a main issue, it appears very difficult to ensure that input files agree among them. Such a heterogeneity produces loss of information during the use of the protocols and generates uncertainty in the final metabolic model. Here we introduce the PADMet-toolbox which allows conciliating genomic and metabolic network information. The toolbox centralizes all this information in a new graph-based format: PADMet (PortAble Database for Metabolism) and provides methods to import, update and export information. For the sake of illustration,

*Intervenant

†Auteur correspondant: marie.chevallier@inria.fr

‡Auteur correspondant: anne.siegel@irisa.fr

the toolbox was used to create a workflow, named AuReMe, aiming to produce high-quality genome-scale metabolic networks and eventually input files to feed most platforms involved in metabolic network analyses. We applied this approach to two exotic organisms and our results evidenced the need of combining approaches and reconciling information to obtain a functional metabolic network to produce biomass.

The main concept underlying the PADMet-toolbox is to provide solutions that ensure the consistency, the internal standardization and the reconciliation of the information used within any workflow that combines several tools involving metabolic networks. In other words, all the information is stored in a single light and human-readable database that can be easily updated. The latter is then used as a single cornerstone which spreads the information all along the workflow. To that goal, the genomic and metabolic information about an organism is depicted using an oriented graph wherein nodes are linked by relations. A PADMet file consists in three parts. Firstly, the "Nodes" part stores information about each node *e.g.* reactions, metabolites, pathways, genes... Then, the "Relations" part depicts all relations between nodes *e.g.* consumption and production connections between metabolites and reactions. Finally, the "Policy" part introduces constraints satisfied by both nodes and relations. For instance, the "consumption" connection necessarily involves 'metabolite' nodes and 'reaction' nodes. This format can be viewed as an extension of both the internal format used by the Biocyc database (*Caspi et al., 2014*), and the SBML format (*Hucka et al. 2003)*, with a very precise definition of fields avoiding the possible confusion generated by SBML fields. It is strongly inspired by RDF with additional flexibility and an easier readability, which is an attractive feature for biologists.

The PADMet-toolbox is wrapped as a Python library that can manipulate and operate the PADMet format. It includes several methods to represent, reconstruct from multiple data sources, analyze and compare metabolic networks. One main procedure consists in creating a reference PADMet from one or several external database such as Metacyc. The latter can be updated, corrected, or enriched with additional nodes and external relationships. The metabolic network to be studied is then represented as a subset of the PADMet-reference database by importing manually created lists of reactions or a SBML file whose identifiers are automatically mapped on the reference PADMet. Based on the latter, several networks can be merged into a combined network while preserving the consistency and the traceability of the data. Moreover, the PADMet-toolbox enables the analysis of a metabolic network, *e.g.* reports about the contents in terms of reactions, metabolites, pathways. The toolbox can also perform the exploration and visualization of data: the whole network as a wiki and subnetworks (*i.e.* pathways) as pictures. Finally, PADMet can be exported to SBML to feed any tool using such format.

In order to illustrate, we used the PADMet-toolbox to implement the AuReMe workflow: AUtomatic REconstruction of MEtabolic networks based on genomic and metabolomic information. The workflow performs high quality metabolic network reconstruction based on sequence annotation and orthology as well as metabolomic data. It includes three main components that can be run in parallel or subsequently. The first part of the reconstruction process consists in the extraction of information from genome annotations. It is run by the Pathway-Tools applications (*Karp et al., 2010*) and output files are converted in PADMet format. The second independent part of the workflow consists in the creation of a metabolic network based on orthology between the studied species and a taxonomically-close template species with a curated metabolic network. The chosen tool to perform this orthologue-based network reconstruction was the Pantograph software (*Loira et al., 2012*), itself based on consensus scores from the InParanoid (*Remm et al., 2001*) and OrthoMCL tools (*Li et al., 2003*). The result of Pantograph is a SBML file that is also converted in PADMet format. Both networks are then merged in one unique and unified network to benefit from the internally standardized annotations of the reference database. The last independent part of the workflow consists in gap-filling a metabolic network (either a network resulting of any of the two previous steps, the merged one or a newly imported one). The gap-filling step aims to complete the network with computed predicted reactions to ensure that a set of metabolic compounds can be produced from the growth media of the studied organism. It

is implemented in the Meneco tool which uses a topological criteria to assess producibility. As above, the toolbox is able to handle inputs and outputs of Meneco and also manual curation performed by the user, who can remove reactions or add particular ones rather than the whole Meneco output. Therefore, the workflow is flexible enough to be iterated both automatically and manually through expert curation until the result of the reconstruction is considered relevant. The PADMet-toolbox enables to check the properties of the network at any time of the reconstruction and facilitates the transition to other metabolic network analysis platforms such as the Matlab Cobra and Raven toolboxes. Both the toolbox and the workflow are available as a Docker image to facilitate their distribution among the scientific community.

The AuReMe workflow was applied to two case-studies. We voluntarily selected species which are distantly related to common model organisms *i.e.* species whose genome or transcriptome annotations cannot deserve a very detailed and specific attention as it was the case for the model species. Therefore, despite the efforts provided for their annotation, genomes of these species contain many genes of unknown function which may generate uncertainties in the network reconstruction process. We considered the cases of an extremophile bacteria (*Acidithiobacillus ferrooxidans* strain Wenelen) and of an eukaryota macro-alga (*Ectocarpus siliculosus*). For each case, the reference PADMet database was fed with the Metacyc 18.5 database content. PADMet tools enabled to unify and compare the results from the various AuReMe reconstruction steps: Pantograph orthology-based network, Pathway-Tools annotation-based network, merging of both, with or without a Meneco gap-filling step. An artificial biomass reaction, based on all the metabolites known to be produced by the organism, was created using PADMet tools to quantitatively simulate the functionality of the organism in FBA, which is a way of assessing the quality of reconstructed metabolic networks.

Along the AuReMe workflow analyses, we noticed that annotations and orthology information was insufficient to make a functional metabolic network, as all the targets are not producible with any individual approach nor the merged network. This leaded to non-production of the biomass and needed to be offset by the Meneco gap-filling step. However, the complementarity of both the annotation-based reconstruction and the orthology based one is an interesting feature. For *A. ferrooxidans*, the resulted network from the merging of Pathway-Tools and Pantograph approaches consisted of 1778 reactions. 44% of them were specific to the first network whereas 23% were specific to the second, illustrating this complementarity and supporting the fact that all reconstruction steps are thus important.

The added-value of the combination of tools in the workflow was also noticeable by comparing the number of unproducible metabolic targets remaining at each step. Merging annotation-based and orthology-based networks increased the number of metabolites topologically producible from the media. Indeed, in the case of *E. siliculosus*, the Pantograph network was only able to produce 2 metabolic compounds out of the 50 experimentally evidenced for this species. The Pathway-tools network produced only 5 metabolic targets. The Pantograph and Pathway-tools merged network was able to produce 27 metabolic compounds. After the Meneco gap-filling step, all compounds were topologically producible. Comparing the reactions added by the gap-filling step to the orthology-based, the annotation-based and the merged networks is an additional way to assess the added-value of the workflow and of the combination of both approaches.The non-merged networks need a greater number of added reactions to produce all the metabolic targets than the merged one. Furthermore, Meneco added a relatively small number of reactions compared to the size of the initial network (less than 5%) but they all were necessary to make the network functional towards the metabolic targets. According to our results, the *E. siliculosus* orthology-based network needed 108 reactions to reach the producibility of metabolic targets. To reach the same producibility, the annotation-based network needed less reactions (66). Considering the resulted network obtained by the merging of both networks, only 42 reactions were necessary to enable the topological producibility of all compounds.

After AuReMe reconstruction, both the bacterial and algal networks were functional, *i.e.* able to produce biomass according to the FBA formalism. Interestingly, performing gap-

filling of *A. ferrooxidans* after each reconstruction step was powerful enough to produce biomass. However, gap-filling the merged networks rather than the individual ones is more relevant as less putative reactions need to be added to make the target compounds producible. Moreover, the analysis of the flux distribution in each network indicated that a larger subnetwork (more routes) can be used to produce biomass in the merged gap-filled network rather than in the two others independently.
Together, our analyses suggest that the combination of tools in the workflow through a local flexible database is a way to reconstruct high-quality metabolic networks. Indeed it lowers the number of reactions that are supported neither by genetic annotation nor by orthology evidence. The uncertainty related to the biological relevance of the reconstruction is thus reduced and less expert-performed manual curation is needed. The AuReMe workflow based on the PADmet toolbox allows handling the heterogeneity of metabolic networks data and preserving the consistency of information in a light and efficient way. It can interact with the various existing *in silico* platforms that aim to reconstruct and/or analyze metabolic networks. In the landscape of numerous metabolic network reconstruction and analyses methods, the PADMet toolbox and AuReMe workflow are flexible solutions to conciliate data and facilitate its processing by biologists and bioinformaticians.

Mots-Clés: data homogenisation, genome, scale metabolic networks, reconstruction workflow, exotic species